https://doi.org/10.2352/ISSN.2470-1173.2018.17.AVM-347

Multiple pedestrian tracking in moving vehicle using online learning of random ferns and feature descriptor of pre-trained shallow Convolutional Neural Networks

SangJun Kim, JaeYeal Nam, and ByoungChul Ko; Dept. of Computer Engineering, Keimyung University, Daegu, S. KOREA*

Abstract

In this paper, we introduce a multi-pedestrian tracking algorithm for tracking from a moving vehicle. The method is based on online learning of a random ferns (RF) tracker model using the output features of a convolutional neural network (CNN). For real-time application in vehicles, an online method is applied within the tracking-by-detection framework where data association between detections and trackers is conducted online. To predict the tracker's position, we perform particle filtering with tracker models inferred from a shallow CNN. In this study, You Only Look Once (YOLO), a real-time object detection system, was adopted as the pre-trained model. Although YOLO has an accurate network for object classification, it is not appropriate for real-time multi-pedestrian tracking. Therefore, we use modified YOLO to obtain a shallow version (S-YOLO) having fewer convolutional layers and fewer filters in these layers. To update the tracker in every frame, positive and negative samples are applied to the S-YOLO and retraining is performed. Then, we extract feature descriptors from the first fully connected layer of S-YOLO to train the RF tracker models. The proposed algorithm was successfully applied to various pedestrian video sequences and yielded a more accurate tracking performance than other existing method.

Introduction

Pedestrian tracking is an essential element of advanced driver assistance systems (ADASs), because it is closely related to the prevention of pedestrian-vehicle collisions. However, pedestrian tracking from a moving vehicle remains very challenging because of the camera's movement, a wide range of deformable pedestrian appearances, cluttered backgrounds, and difficult real-time constraints [1].

There are two types of object tracking approaches. The first, deterministic optimization, seeks the optimal solution to link observations from each of the frames in the image sequences based on offline global optimization of all object trajectories. Starting from the output of simple object detectors, this approach builds a network graph in which every node is an observation fully connected to future and past observations. Although this approach usually achieves a better tracking performance than online linking methods, it is not suitable for the task of real-time online tracking because of its considerable computational complexity. The representatives of this approach are Bipartite Graph Matching, Dynamic Programming, and Min-cost Max-flow Network Flow. The second approach type, probabilistic inference, estimates the probabilistic distribution of the target status (size, position, velocity) by using a variety of probability reasoning methods based on existing observations. This approach requires only the existing observations. Therefore, it is appropriate for real-time tracking based on online learning. The representatives of this approach are the Kalman filter, extended Kalman filter, and particle filter.

Recently, in many algorithms for tracking systems, convolutional neural networks (CNNs) have been applied, because CNNs learn feature representations as part of their training and outperform heuristic, hand-crafted features in several vision problems. Li et al. [2] proposed a target-specific CNN for object tracking; the CNN is re-trained incrementally during tracking with new examples obtained online. This approach tracks only one object and uses a shallow CNN for online learning. However, because online learning of CNNs incurs high-level computational complexity, the use of multiple objects tracking (MOT) that requires an individual tracker model is not feasible in real time.

Many tracking approaches based on CNNs focus on data association using a network trained offline to determine whether two detections belong to the same trajectory [3][4], because online learning based on CNNs is not straightforward owing to the large network size and lack of training data. However, data association based on a network trained offline is a challenging task in the presence of occlusions, missing objects, and false alarms. Therefore, this study introduces real-time data association for multi-pedestrian tracking with a pre-trained CNN and online tracker learning over two consecutive frames.

The objective of this study was to design a system, based on online learning of a tracker model using the output features of a CNN, for tracking multi-pedestrians from a moving vehicle in real time that is effective regardless of the camera type, where a color camera is used for daytime and a thermal camera for nighttime. In our system using a target-free CNN with a tracker model for pedestrian tracking, the tracker model is re-trained incrementally during tracking with new examples and their features obtained from the output features of the CNN. This approach tracks multi-pedestrians and uses a shallow CNN network for online learning.

Multi Object Tracking Based on Shallow CNN and Random Ferns

Online tracker learning based on and shallow CNN and random ferns

For real-time application in vehicles, an online method based on probabilistic inference is applied within the tracking-by-detection framework, where data association between detections and trackers is conducted online. To predict the tracker's position, we perform

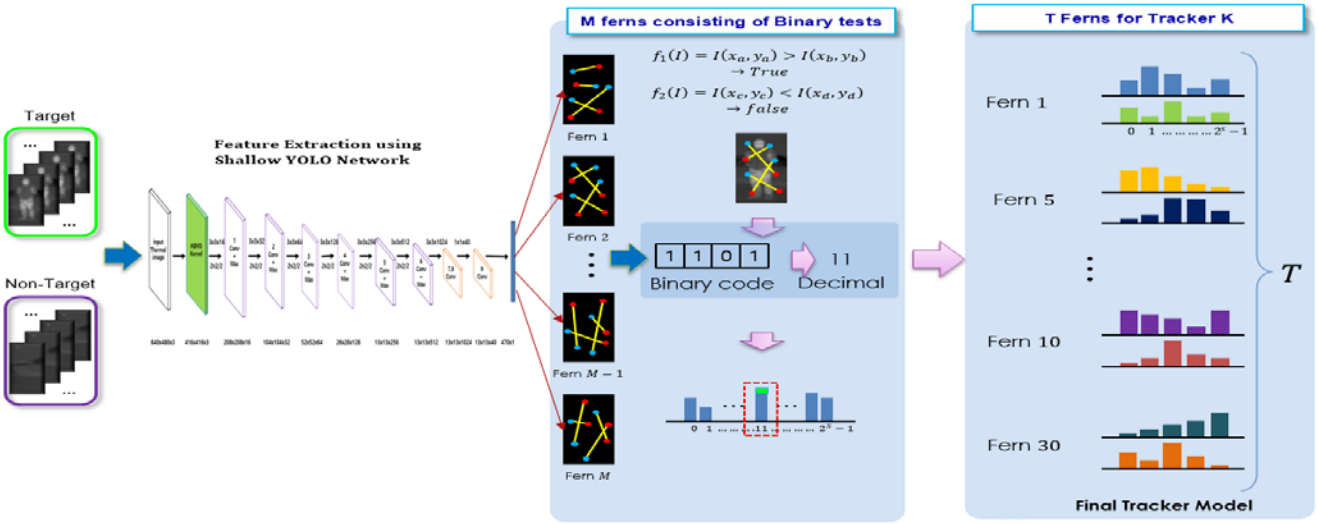

Fig. 1 Overall procedures of tracker model retraining. The output features of the shallow YOLO are extracted and the desired M ferns consisting of binary tests are learned as a tracker model.

particle filtering with tracker models inferred from a CNN. The CNN is pre-trained on a large number of images. In this study, YOLO [5], a real-time object detection system, was adopted as the pre-trained model. YOLO [5] uses a single neural network to predict the bounding boxes and class probabilities directly from full images in a single evaluation. Because the entire detection pipeline is a single network, the detection performance end-to-end can be quickly and directly optimized.

A YOLO network consists of 24 convolutional layers followed by two fully connected layers, and uses 1 × 1 reduction layers followed by 3 × 3 convolutional layers. Although YOLO has a fast network for object detection, it is not appropriate for real-time multi-pedestrian tracking. Therefore, we modified YOLO to obtain a shallow version having fewer convolutional layers and fewer filters in these layers. This shallow version of YOLO (S-YOLO) consists of nine convolutional layers and six max pooling layers followed by one fully connected layers. For training the S-YOLO, we use pre-trained convolutional weights that are trained on ImageNet as the initial parameters of the system [6]. To update the tracker, positive training examples are sampled from the bounding box of the associated detection. The negative training set is sampled from nearby targets and random samplings from background and other pedestrian regions. The training samples are applied to the pre-trained S-YOLO and we extract feature descriptors from the first fully connected layer, because it tends to capture the general characteristics of pedestrians and has shown an excellent generalization performance in many other applications.

From the feature descriptors extracted through the S-YOLO, we trained a tracker model. The tracker model fills the important role of estimating the most likely location among several particles and handles appearance changes and the drifting problem. In this study, we used random ferns (RF), which has been applied in many computer vision studies, such as object detection, key-point recognition, and object tracking. Ferns model the dependencies between intra-group features while preserving the assumption that inter-group features are conditionally independent. RF has a fast run-time performance and is robust, because it does not perform evaluations of the binary tests that comprise the ferns [1].

Figure 1 shows the overall procedures of the proposed S-YOLO-based online tracking model. After a pedestrian has been tracked, positive samples are collected near the tracker location and negative samples from background and other pedestrians. From the training data, the output features of the S-YOLO are extracted and the desired M ferns consisting of binary tests are learned as a tracker model.

Multi object association checking

Recently, tracking approaches were proposed that seek the optimal solution to link observations from each of the frames in the image sequences. This approach type is suitable for the task of offline tracking, because observations from all the frames, or at least a time window, are used. However, to link detection responses to pedestrians' trajectories in a real driving situation, we apply the simple and fast Hungarian algorithm using a short-term tracklet. First, we use a particle filter that recursively approximates the posterior distribution using a finite set of weighted samples and estimate the observation likelihood for the particle weights, for a particle p of tracker tr, using a trained individual tracker based on RF with the S-YOLO output. After a new tracker location is detected, to evaluate the matching score for each tracker-detection pair (tr, d), we linearly combine the observation likelihood $pr(d|tr)$ using RF, location distance $Dist(tr, d)$, and overlap ratio $OR(tr, d)$ between tracker tr and detection d. The data association problem is defined by a linear program with the objective function

$$S^* = \arg\max_{d} \left\{ a \cdot pr(d|tr) + \beta \cdot {}^{1}/_{Dist(tr,d)} + \gamma \cdot OR(tr,d) \right\} \quad (1)$$

where α, β, and γ are parameters that can be adjusted according to the applications.

Finally, the trackers and their RFs are retrained based on the feature vectors of the S-YOLO output according to the matching results, and the tracker's state is updated by combining the states of the current tracker and the detection information.

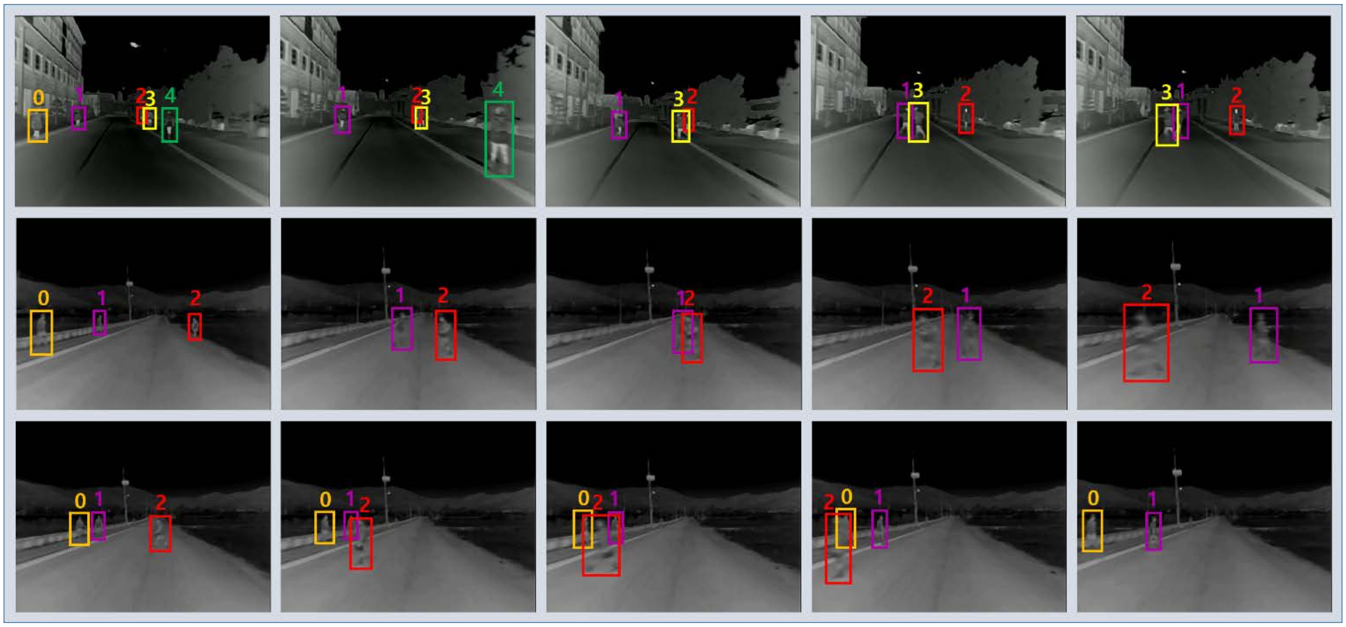

Fig. 2. Tracking results of the proposed method in a moving car. The marked number is the tracker's ID and it has the same color if a tracker is the same in subsequent frames

Experimental Results

To evaluate the tracking performance of the proposed system, we used one thermal video dataset, the KMUTD dataset [1], which consists of 10 video sequences captured from a thermal camera. The total amount of test data for winter three video sequences constitutes 2,848 frames, and the total amount for two summer video sequences constitutes 1,153 frames.

The tracking algorithm based on the shallow YOLO with online RF learning shows a higher performance level, because S-YOLO extracted optimal feature including specific characteristics of individual pedestrian and the tracker model was updated based on S-YOLO feature in the process of online learning. To reduce the uncertainty of the performance evaluation, we computed the average Multiple Object Tracking Precision (MOTP) for the five KMUTD dataset.

To evaluate the performance of the proposed algorithm, we compared its performance with that of recent state-of-the-art methods, Kwak et al.'s [1] approach that tracks multiple pedestrians using an online RF with conventional handcrafted feature. Figure 3 shows the MOTP comparison between two methods for all video sequences.

According to our experiment, the average MOTP of the proposed algorithm are 92.7%. The average MOTP of the proposed approach is significantly improved than that of comparison method about 9.1%. The main reason of improved the performance is that S-YOLO detected pedestrian more correctly than conventional approach of [1] and S-YOLO was re-trained every frame to extract pedestrian oriented feature.

All the experiments were conducted on an Intel Core i7 PC with 8 GB of RAM running Microsoft Windows 7. For the graphical processing unit (GPU), we used an NVIDIA GTX 1080. All experiments were conducted based on the GPU and Compute Unified Device Architecture (CUDA) for deep learning. The computational speed of the detection method was approximately average 34 frames per second. However, the computational speed of the tracking including detection, online RF learning, and association time was approximately 2.0 frame per second. In the second version of our system, however, we expect to reduce the computational speed by adjusting the structure of the S-YOLO without degrading the tracking performance.

Conclusion

The main contributions and overall procedures of our study can be summarized as follows; 1) we introduced a multi-pedestrian tracking algorithm based on online learning of a RF tracker model using output features of S-YOLO in a moving vehicle in real time. 2) we modified YOLO to obtain a shallow version having fewer convolutional layers and fewer filters in these layers. 3) to update the tracker in every frame, positive and negative samples were applied to a S-YOLO and performed the retraining. Then, we extracted feature descriptors from the first fully connected layer of S-YOLO to train RF tracker models. 4) in our target-free S-YOLO with tracker model for pedestrian tracking, the tracker model was

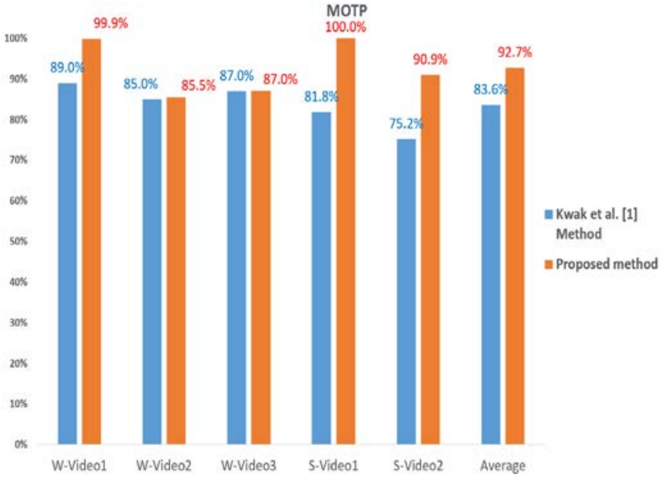

Fig. 3. MOTP performance comparison of the two methods about five thermal video sequences

re-trained incrementally during tracking with new examples and their features obtained from the output features of the YOLO.

In the future, we plan to solve computational overload related to feature extraction of S-YOLO and online learning of RF by reducing the layer structure and designing light version of CNN.

Acknowledgement

This research was supported by Basic Science Research Program through the National Re-search Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1A09916581) and partially supported by Future Vehicle Leading Technology Development Project funded by Daegu City (DG-2017-01).

References

[1] J. Y. Kwak, B. C. Ko, and J. Y. Nam, “Pedestrian tracking using online boosted Random Ferns learning in far infrared imagery for safe driving at night," IEEE Trans. Intell. Trans. Sys., vol. 18, issue 1, pp. 69-81, Jan., 2017

[2] H. Li, Y. Li, and F. Porikli, “Deeptrack; learning discriminative feature representations by convolutional neural networks for visual tracking,” in British Machine Vision Conference, Nottingham, U.K., Sept. 2014.

[3] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, “Learning by tracking: Siamese CNN for robust target association,” in IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017.

[4] J. Son, M. Baek, M. Cho, and B. Han, “Multi-object tracking with quadruplet convolutional neural networks,” in IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017.

[5] J. Redmon, S. Divvala, S. Girshick, and A. Farhadi, “You only look once: unified, real-time object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[6] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017.

Author Biography

Sang Jun Kim received his B.S. degrees in Computer Engineering from Keimyung University, Daegu, Korea, in Feb. 2017. He is currently a M.S. student in the Department of Computer Engineering, Keimyung University, Daegu, Korea. He received the best paper award in 2017 from Korea Computer Congress (KCC2017). His current research interests include advanced driver assistance systems using computer vision and deep learning (tmsor5@naver.com).

JaeYeal Nam received the B.S. and M.S. degrees in electronics engineering from Kyungpook National University, Daegu, Korea, in 1983 and 1985, respectively. And He received Ph.D. degree in electrical engineering from University of Texas at Arlington in 1991. His research interests are in the areas of image and video compression, contents-based medical image retrieval and vision-based fire detection algorithms. He is currently vice present of Keimyung University from 2014 (jynam@kmu.ac.kr)..

Byoung Chul Ko (Corresponding author) received his B.S. degree from Kyonggi University, Suwon, Korea, in 1998 and M.S. and Ph.D. degrees in Computer Science from Yonsei University, Seoul, Korea, in 2000 and 2004, respectively. He is currently a professor in the Department of Computer Engineering, Keimyung University, Daegu, Korea. His current research interests include content-based image retrieval, vision-based fire detection, advance driver assistance systems, and facial emotional recognition (niceko@kmu.ac.kr).